

SPEECH VARIATION AND THE USE OF DISTANCE METRICS ON THE ARTICULATORY FEATURE SPACE

Louis ten Bosch

CSLT, Radboud University, Nijmegen, the Netherlands

l.tenbosch@let.ru.nl

ABSTRACT

This paper describes ongoing research on the relation between variation in speech in the articulatory-acoustic domain and the variation as represented in the symbolic domain. More specifically, we address variation in speech as represented by articulatory features, and the description of variation in phone annotation and segmentation. Variation in speech is quantified by using distance metrics defined on the space spanned by articulatory features. We will show a very good correspondence between locations of events in the articulatory feature trajectories on the one hand, and the phone boundary locations as defined by manual segmentation on the other. This indicates that the asynchronous articulatory representation at least captures the information in the segmentation on phone level.

The proposed technique can be used for designing alternative representations of the speech signal to describe phonetic-linguistic phenomena, including intrinsic variation, and for automatic annotation and segmentation procedures.

1. INTRODUCTION

Mainstream approaches in automatic speech recognition (ASR) systems assume that the speech signal can be represented as a sequence of discrete symbols (e.g. phone-like symbols). This ‘beads-on-a-string’ paradigm ([15]), based on the discrete symbolic representation of phonemes (e.g. [23]), makes it difficult to adequately variation in speech. Variation that is due to pronunciation variation, speaking styles, inter-speaker differences, accents etc. primarily takes place in a continuous domain, often with effects on the sub-phonemic level, rather than in a discrete domain. The description of variations in a continuous domain by using discrete symbolic representations is not necessarily inadequate, but the relation between variations in the continuous speech domain and the discrete representation is a result of compromises (cf. [1][8][9][17]). The limitations of the present mainstream approach in ASR have inspired many researchers to argue that fundamentally better computational paradigms for speech processing can be achieved by modelling the underlying processes of speech *production*, such as co-articulation and assimilation, rather

than modelling the surface effects on the resulting acoustic speech signal. In this area, progress has been made by using specifically trained articulatory feature classifiers ([6][10][11][13]16][19]), also used in the context of event detection ([12]). In these studies, the choice of the specific set of articulatory features is largely inspired by both the theory of distinctive features ([4]), in which phonemes are represented in terms of binary features (for example voiced/voiceless, rounded/unrounded), and by the gestural theory of speech production ([3]), according to which features may be multi-valued (e.g. the feature *height* might attain the three values *high*, *mid* or *low*).

In this study, we apply an articulatory feature representation using a feature set including manner of articulation, place of articulation, voicing, front-back and rounding, each of these features having a cardinality between 3 and 8. This choice is in line with [6] and [20]. The combination of all articulatory features comprises in total 28 (i.e. the sum of all cardinalities) continuous-valued functions over time, which results in a sequence of articulatory feature (AF) vectors (one each 10 ms). The output of the articulatory classifiers yields a gradual, rich and redundant representation that is articulatory motivated, directly estimated from the signal.

The AF representation of speech offers several advantages compared to the usual symbolic description (phone sequences). Firstly, dependent on the exact training methods for the feature classifiers, AFs provide a description of the speech signal allowing loose synchrony between articulatory features, in contrast with linear phone representations which explicitly impose strictly synchronous feature transitions. AF representations are assumed to be principally better able to overcome limitations imposed by the beads-on-a-string paradigm, since certain types of speech variation often boil down to asynchrony between ballistic movements of articulators (such as gradual incomplete nasalization of vowels followed by nasals, word-final partial devoicing, and schwa epenthesis in consonant clusters).

Second, there is a growing number of indications that fine phonetic (sub-phonemic) details are indeed relevant for human speech processing; human lexical decoding can be shown to be mediated by subphonemic details (e.g. [22]). The appropriateness and validity of any computational model for the human processing of speech should therefore

depend on the capability to effectively handle these subphonemic details. Here, the term *detail* may cover clearly noticeable allophonic variation, and is not restricted to phenomena that are small in the sense of ‘close to psychoacoustic thresholds’ ([22]).

The aim of this paper is to show relations between articulatory feature representations and phonetic events, by applying distance metrics on the articulatory space to quantify the amount of observed variation in the speech signal. Here, the term *event* relates to a salient, sudden event in the output stream of the AF classifiers interpreted as a time-varying high-dimensional signal.

One of the approaches to impose a structure on the bottom-up derived AF streams is based on the application of Dynamic Bayesian Networks (DBN, [2][21]). In this paper, we will follow another approach and restrict ourselves to investigating the structure that emerges from the ‘raw’ output of the AF classifiers themselves. As an example of the power of such a structural analysis, we will relate the variation in the signal (as represented by variation in AFs) with the information in hand-crafted segmentations as provided after human phone-level annotation. AFs, endowed with a proper metric, appear to provide information about the location of hand-crafted inter-phone boundaries and therefore are a promising alternative for more sophisticated event-parsing approaches (e.g. [7]). This result shows that the information in asynchronous representations is at least as rich as the temporal information in the phone-level ‘beads-on-a-string’ segmentation.

This paper is organised as follows. The next section is devoted to a brief introduction to the design and training of the AF classifiers. The third section describes the database of spontaneous speech that was used in this study, while the fourth section discusses the method, the analyses and the results. The final section concludes with a discussion and remarks for further research.

2. ARTICULATORY FEATURE CLASSIFIERS

The articulatory feature values that are exploited in this study are obtained by Artificial Neural Nets (ANN). Each of the six features (manner, place, front-back, voicing, rounding, and ‘static’, see table I) is represented by one ANN. Each ANN is trained on ‘canonical’ feature transcriptions that are obtained by combining the (manually obtained) phoneme transcription of the speech signal and a phone-to-feature table. The method that was used is basically the same method as applied elsewhere (e.g. [6]). For the ANNs used in this paper, we applied the NICO-toolkit ([18]).

In parallel, the six AF classifiers provide information without imposed structure: the strict dependency of the features observed on the canonical training samples is lost due to independency between the classifiers, and so on a test

set AF output vectors may deviate from the ‘canonical’ 0/1 AF vectors that belong to canonical phones themselves. As explained earlier, the AF output consists of 28 parallel signals, each signal updated every 10 ms, with ‘fuzzy’ output values between 0 and 1.

In the context of this paper, a number of observations are relevant with respect to the training and test procedure of ANNs.

- 1) The training is based on phone-based articulatory feature descriptions and all features are presented synchronously. Since all classifiers are trained independently of each other, the output of the classifiers on a test utterance may be (and usually is) asynchronous. The larger this asynchrony, the more non-canonical AF combinations will occur.
- 2) The deviation between observed ANN output and canonical AF vectors (e.g. as measured by the covariance of the set of AF vectors associated to one phone) is based on feature asynchrony and on the fuzziness of the ANN output. While the fuzziness of the ANN can be alleviated by a winner-takes-all post-processing, the remaining contribution from asynchrony is related to intrinsic variation in the speech signal.
- 3) The overall complexity of 6 parallel classifiers (measured in terms of number of node-node connections) is much lower than the complexity of one monolithic classifier. [11] has shown that the division of the overall problem into specialised subproblems leads to improved robustness to noise.

Table I The six features with the 28 values used in this study.

<i>Feature</i> (card)	Values
<i>manner</i> (6)	approximant, fricative, nasal, stop, vowel, silence
<i>place</i> (8)	(labio)dental, alveolar, velar, high, mid, low, silence
<i>voicing</i> (3)	voiced, voiceless, silence
<i>rounding</i> (4)	rounded, unrounded, nil, silence
<i>front-back</i> (5)	front, central, back, nil, silence
<i>static</i> (3)	static, dynamic, silence

3. DATABASE DESCRIPTION

In this study, we have used the IFACorpus ([21]), a database of spoken Dutch. It contains recordings of 4 male and 4 female speakers, varying from 15 to 66 years in age. For all utterances, manually corrected labelling and segmentation on phone and word level are available. Metadata include education level, birth place, and smoking habit and contain more information than is available in the much larger

Spoken Dutch Corpus (CGN, [14]). The transliteration of the IFACorpus is according to the CGN-protocol. Compared to CGN, the amount of speech per speaker is much larger (40 min/speaker) and more speaking styles have been recorded (8, varying from spontaneous monologues to read-aloud word lists). A number of 19867 utterances have been transcribed (a bit more than 5 hours).

Two subcorpora (retold stories in the form of long monologues, and randomly presented sentences, in total about 140 minutes of speech) have been selected for this study. The total number of utterances is 2650. All speech material has been converted to 16 kHz 16 bits/sample mono wav files. The phone alphabet was cleaned up to contain 50 different phones – apart from the basic phones, the IFACorpus also uses palatalised variants. There is one silence symbol.

The selected subcorpus was divided into a training set (1978 utterances), a ‘validation set’ (100) and a test set (572 utt; 44m10s). The test set consisted of the speech from one male and one female who were kept separate, while speech from the other 6 speakers was used for training and validation.

4. ANALYSIS METHOD AND RESULTS

4.1. Analysis method

The training and validation set have been applied for the training of the six different ANNs that model the 6 features (manner, place etc.). Table II, second column, shows the classification results on the IFACorpus test set (the accuracy of the individual classifiers on frame level in percentage correct). For the sake of completeness, we added the ANN results obtained on the TIMIT test set after training on the TIMIT training set, but since transcription methods and database specifications differ in detail, a further cross-database comparison hardly makes sense.

After training, the classifiers were used to produce AF vector sequences for all test data, overall resulting in about 265000 vectors of dimension 28.

Table II Frame-based accuracy of feature classifiers (in perc.) on the IFACorpus and TIMIT test set.

Feature/ANN	IFACorpus	TIMIT
<i>manner</i> (6)	84.7	86.5
<i>place</i> (8)	76.7	78.6
<i>voicing</i> (3)	93.5	92.0
<i>rounding</i> (4)	87.4	86.0
<i>front-back</i> (5)	83.6	83.0
<i>static</i> (3)	89.7	81.0

In order to align the AF vector sequences with the manual phone labels, each utterance was associated with a matrix in

which both the AF information as well as the encoded segment label information was stored. For all test utterances concatenated, this results in a matrix of dimension 265000 x 29.

Each utterance is represented by a trajectory in AF space. In the literature, several methods have been investigated to define events on the basis of the sequence of AF vectors and to relate these events with the canonical symbolic information. Using a parsing technique similar to HMM decoding based on phonotactic information, Hacıoglu and colleagues [7] segmented the vector stream into ‘events’ that could be linked to phonetic segmentation.

In order to relate AF variation with symbolic variation, we conducted an analysis in two steps. Firstly, we analyzed the variation as evidenced in the set of AF vectors themselves, given the set of canonical AF vectors (these canonical vectors one-to-one corresponding to the 50 phones). To that end, we investigated a set of possible metrics for defining ‘distance’ in AF space. Second, the variation was investigated along the time dimension and related to the information by the manual segmentation.

Step 1: metrics

The first step is based on the assumption that variation in speech is adequately reflected by variations in AF space, and that this variation can be quantified by an appropriate metric. To that end, we pre-selected three different metrics that are widely used.

- a) The Euclidean metric and its weighted version:

$$D^2 = \sum \alpha_i (v_i^{(1)} - v_i^{(2)})^2$$

- b) The cosine-based metric:

$$A = (v^{(1)}, v^{(2)}) / (|v^{(1)}| |v^{(2)}|);$$

$$D = \arccos(A)$$

- c) The weighted city-block distance:

$$D = \sum \alpha_i |v_i^{(1)} - v_i^{(2)}|$$

In these expressions, $v^{(1)}$ and $v^{(2)}$ denote two AF vectors and the summations run over all components. As the weightings $\{\alpha_i\}$ determine the contribution of individual component differences to the overall distance, they directly relate to articulatory compensation and other phenomena of cue trading.

The metrics quantify variation in AF space in a different way. Figure 1 indicates how the Euclidean distance, the cosine distance and the city-block distance relate to one another. The distances are evaluated using equal weighting for all components of AF vectors with dimension 28, and

were left unscaled before plotting. The distances clearly do not relate in a linear fashion, and therefore deal differently with small and large differences.

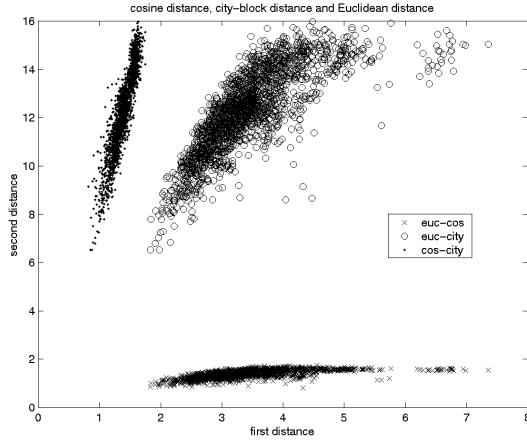


Figure 1. The cosine distance, the city-block distance and the Euclidean distance plotted against one another.

Among the three metrics, the cosine distance is the single distance with an additional intrinsic property that makes it particularly useful in AF space. It normalizes the vectors in such a way that differences due to their individual (Euclidean) length become irrelevant – a property that the other two distances do not have. While for the Euclidean and city-block metric the vector $(0, 0, 0.6, 0.6)$ is different from $(0, 0, 0.5, 0.5)$, these vectors have distance 0 according to the cosine distance. Since the AF classifiers do not normalize their output, the cosine distance removes some of the unnecessary freedom by discarding length. In the sequel, we have employed the cosine distance on the basis of this additional normalization property.

Step 2: speech variation in terms of local distances

The AF trajectory that is produced by the AF classifiers can be interpreted as a vector sequence

$$\{\dots, v_{n-1}, v_n, v_{n+1}, \dots\}$$

The vectors are not augmented in the sense of containing delta and delta² information. Variation over time is therefore encoded in distance sequences such as

$$\{\dots, D(v_{n-1}, v_{n-2}), D(v_n, v_{n-1}), D(v_{n+1}, v_n), \dots\}$$

and similar for higher lags.

Since AF detectors are trained to yield canonical AF vectors corresponding to the frame-assigned phone label, the output is likely to be constant (i.e. $D(v_n, v_{n-1})$ small, approx. 0) for a range of frames in the ‘stable’ part of a phone segment. This suggests that the parts during which the distance between

consecutive frames peaks may carry relevant cues for event segmentation. That this holds to a large extent is shown in the next subsection.

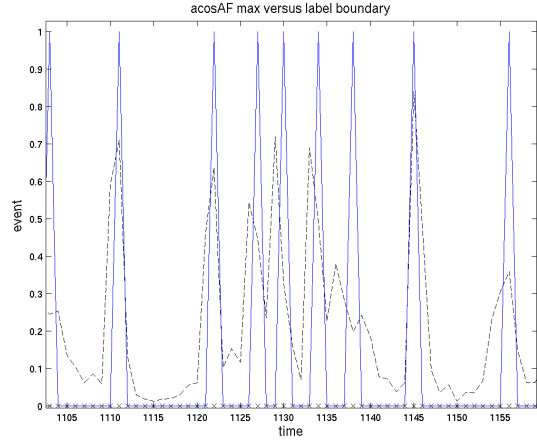


Figure 2. Example of alignment between cosine distance peaks and segment boundaries. For an explanation see the text.

4.1. Results

Figure 2 shows a random chosen snapshot of about 50 frames (spaced 10 ms) in the middle of a Dutch utterance from the test set in the IFAcopus. In the figure, two plots are plotted against time (horizontal axis). First, the dashed function with a range between 0 and 0.7 shows the cosine-based distance between consecutive pairs of AF frames. The information given by this curve is broadly comparable to (but in detail different from) the information in mainstream-ASR delta features, as it gives an overall ‘summary’ of changes in the entire delta-components. Secondly, the peaks of the triangle-shaped contour (range 0-1) show the location of the boundaries between the phone labels according to the manual segmentation. In the figure, a clear relation is visible between the locations of the maxima in the cosine-distance and the boundaries in the manually corrected segmentation (triangle peaks). The figure suggests that segment boundaries are associated with local *speed* along the AF trajectory. The quality of this alignment was investigated on a 10-min selection of the entire test set, by making a histogram of the distances (in frames) between a cosine-distance maximum and the *closest* segment boundary. Of the found 6810 cosine-maxima in total, 215 could not be associated with a segment boundary within the range $[-5, 5]$. Of the remaining 6575 maxima, 40 percent *coincide* with the manual boundary, while 89 percent could be assigned a boundary within 25 ms from the cosine peak. This high alignment score is a strong indication for the applicability of cosine-distance for data-driven event parsing in AF space. Evidently, this finding does not imply that all phone boundaries are assigned to a cosine-distance peak, which is

plausible since the AF-based event detection is based on a combination of bottom-up information and the information about the manual segmentation that is implicitly encoded in the ANNs. The value of 89 percent within 25 ms is comparable to the accuracy of 84 percent within 20 ms (reported in [24], table 5) for the position of phone boundaries by automatic segmentation.

It is interesting to describe the differences between the acoustic event detection (cosine distance peaks) and manual segment boundary locations in terms of articulatory properties. Overall, there is a slight tendency (of 0.2 frames) for the cosine distance peaks to be too early, which might be attributable to certain initial settings in the signal processing. When each peak-boundary assignment is investigated by using the articulatory features of the involved neighbouring phones, a number of phone transitions tend to stand out; these are mentioned in table III. In this table, a negative distance average (avg.) indicates that the manual segment boundary *precedes* the cosine distance peak (and v.v.). The number of occurrences per transition varies from 51 to several hundreds. Only the transitions with extreme negative and positive distances are shown. Paired *t*-tests indicated that cases marked with [*] are significantly different ($p < 0.05$) from the fricative-fricative transition (first line).

Table III Average (in number of frames) and standard deviation of the difference (diff.) between cosine-peak location and manual boundary. Only the transitions with extreme negative and positive distances are shown

Manner features	avg. (st.dev.)
Fricative-fricative	-0.57 (1.6)
Vowel-vowel	-0.31 (1.8)
....	
Silence-approximant	0.49 (1.8) [*]
Approx.-stop	0.63 (1.6) [*]
Vowel-silence	0.64 (2.1) [*]
Nasal-approx	0.66 (1.0) [*]

5. DISCUSSION

5.1. Variation in speech

A substantial amount of research has been devoted to unravelling the major components of variation in speech. Here, we address variation by specifically looking at events in the time-evolution of the AF vector sequence. Three metrics were considered to measure variations in AF space. In the paper, we have opted for using the simplest metric that also incorporates a normalization (cosine distance), and we have shown that the location of peaks in the cosine distance can be aligned to a large extent with the boundary locations in the manual phone segmentation. This method works remarkably well for carefully pronounced read-aloud speech. Manual segmentations are known to be accurate up

to two to three frames, and much less reliable in case of transitions between phones with high phonetic similarity. It is therefore challenging to interpret the method as a novel way to parse the AF feature stream into events that are defined intrinsically, that is, by the use of metrics defined on the AF space, without external labelling or segmentation.

It seems that the event parsing along AF trajectories as proposed in this paper is a step back from the claimed advantage of asynchronous speech representations. This is not true. On the contrary, the fact that the majority of phone-based segmentations can be detected on the basis of asynchronous feature representations in combination with an appropriate metric shows that the chosen representation at least contains the information encoded in segmentations on phone level, and is likely to be much richer than the beads-on-a-string representation.

Manual phone-level segmentations were obtained by the combination of auditory and visual inspection. That means that the information in the segmentation is determined by more knowledge bases than auditory channel. Nevertheless, the AF classifiers appear able to model this segmentation process by only using information that was extractable from the signal. Discrepancies between model and manual segmentations may partly be attributable to this difference.

5.2. Current shortcomings and near-future plans

Given the current results, it is an open question to what extent the type of distance measures distinguishes fine detail in the alignment with manual segmentation. For shorter distances close to 0, all metrics will provide about the same result, but the metrics deviate for larger distances, thereby putting more weight to different types of distinctions.. This means that event parsing along the AF trajectory may result into essentially different segmentations along the trajectory for different metrics. This topic is currently under investigation but precise relations are still unknown. Also unclear is the cue trading (by using weights) and the precise quantification of asynchrony. The variation of observed AF vectors around a canonical AF vector is the combined contribution of both the feature asynchrony and the statistical variation in the classifier output, and we plan to explicitly unravel these effects in the near future.

It is tempting to exploit the phenomena described here in terms of design principles for alternative procedures for data-driven annotation and unit selection.

6. CONCLUSION

We have shown that by using an articulatory feature (AF) representation in combination with an appropriate distance function, a segmentation of the speech signal can be obtained with a high agreement (89 percent within 20 ms) with hand-crafted phone-segmentations. The automatic segmentation is based on measuring intrinsic variation using

a local distance function along the speech trajectory in AF space.

7. ACKNOWLEDGEMENTS

Thanks to Lou Boves, Annika Hämäläinen, Christophe Van Bael and anonymous reviewers for comments, to Mirjam Wester for assistance with the ANN set-up, and to Rob van Son for assistance in making the IFACorpus available.

8. REFERENCES

- [1] Bael, C. van, Heuvel, H.v.d., Strik, H. (2004). Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora. Proceedings of Interspeech (ICSLP) 2004, Jeju, Korea (cd-rom).
- [2] Bilmes, J., (2002). GMTK: The Graphical Models Toolkit. URL <http://ssli.ee.washington.edu/~bilmes/gmtk/>
- [3] Browman, C., Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica* 49, pp. 155–180.
- [4] Chomsky, N., Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York, NY.
- [5] Frankel, J. (2003). Linear dynamic models for automatic speech recognition. Ph.D. thesis, The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK.
- [6] Frankel, J., Wester, M., King, S. (2004). Articulatory feature recognition using dynamic Bayesian networks. Proceedings of Interspeech (ICSLP) 2004, Jeju, Korea, (cd-rom).
- [7] Hacioglu, K., Pellom, B., Ward, W. (2004). Parsing speech into articulatory events. In: Proceedings of ICASSP '04. Montreal (cd-rom).
- [8] Hämäläinen, A., de Veth, J., Boves, L. (2005). Longer-Length Acoustic Units for Continuous Speech Recognition. Proceedings EUSIPCO, Antalya, Turkey (cd-rom).
- [9] Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., Sen, Z., (2001). What kind of pronunciation variation is hard for triphones to model? In: Proceedings of ICASSP. Vol. 1. pp. 577–580.
- [10] King, S., Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language* 14 (4), pp. 333–353.
- [11] Kirchhoff, K. (1999). Robust speech recognition using articulatory information. Ph.D. thesis, University of Bielefeld, Bielefeld, Germany.
- [12] Li, J, and Lee, C.H.. (2005). On designing and evaluating speech event detectors. Proceedings Interspeech-2005 (cd-rom).
- [13] Metze, F., Waibel, A. (2002). A flexible stream architecture for ASR using articulatory features. In: Proceedings of ICSLP, Denver, CO, (cd-rom).
- [14] Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., Smith A. (Eds) *New Frontiers of Corpus Research* (pp. 105-112). Amsterdam: Rodopi.
- [15] Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string' model of speech. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. Vol. 1. Keystone, Colorado, USA, pp. 79–83.
- [16] Richards, H. B., Bridle, J. S., (1999). The HDM: A segmental hidden dynamic model of coarticulation. In: Proceedings of ICASSP. Vol. 1. Phoenix, AZ, pp. 357–360.
- [17] Strik, H., Cucchiari, C. (1999). Modeling pronunciation variation for ASR: a survey of the literature. Special Issue of Speech Communication on 'Modeling Pronunciation Variation for Automatic Speech Recognition', Vol. 29, No. 2-4, pp. 225-246
- [18] Strom, N. (1997). Phoneme probability estimation with dynamic sparsely connected artificial neural networks. The Free Speech Journal Issue #5.
- [19] Wester, M., Greenberg, S., Chang, S. (2001). A Dutch treatment of an elitist approach to articulatory-acoustic feature classification. In: Proceedings of Eurospeech, Aalborg, Denmark, pp. 1729–1732.
- [20] Wester, M. (2003). Syllable classification using articulatory-acoustic features. In: Proceedings of Eurospeech, Geneva, Switzerland (cd-rom).
- [21] Wester, M., Frankel, J., King, S. (2004). Asynchronous articulatory feature recognition using dynamic Bayesian networks. In: Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop. Vol. 104. Kyoto, Japan, pp. 37–42 (SP2004-81-95).
- [21] Son, R.J.J.H. van, Binnenpoorte, D., Heuvel, H. van den, Pols, L. (2001). The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database, Proceedings of Eurospeech, Aalborg, Denmark, Vol. 3, pp. 2051-2054.
- [22] Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding, *Journal of Phonetics*, 31, pp. 373-405.
- [23] Trubetzkoy, N. (1939). Grundzüge der Phonologie (Principles of Phonology). Travaux du Cercle linguistique de Prague 7.
- [24] Wesenick, M.-B., Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. Proceedings ICSLP, Pittsburgh (cd-rom).